晓宇：亿级用户个性化品类推荐实战解析

2018年4月16日，周一晚上8点30分，美团平台金刚区个性化推荐算法和业务研发主要负责人晓宇带来了主题为《亿级用户个性化品类推荐实战》的交流。以下是主持人天怡整理的问答实录，记录了作者和读者间问答的精彩时刻。

内容提要：

- LR+FM 混合模型具体怎么做？
- 数据清洗采用的是什么技术？用过 ETL 工具吗？
- Spark 跟 Hadoop 的区别是什么？为什么要使用 Spark？
- FM 中的 k 表示什么？
- Spark 在贵公司的实际应用场景有哪些？Sparkmlib 是否比 skitlearn 和 TensorFlow 更实用一些？
- Java Web 程序员如何转型推荐系统算法工程师？
- 请问为何不使用深度学习来推荐？如何做的特征筛选？FM 是 sparkmlib 自带的包还是自己实现的？以及用户画像是怎么实现的？
- 推荐系统有做 abtest 吗？
- 请问模型是实时训练的吗？
- 设备 ID 到用户 ID 转换，扩展未登录场景，请问怎么扩展到未登录场景，比如 H5 未必存在获取到设备信息？怎么设计推荐评测系统进行验证，以便不断优化推荐算法？
- 美团推荐的召回是怎么做的，精排只用了 FM+LR 吗，重排是怎么做的？
- 美图推荐的用户画像有多少个特征，如果特征很大，怎么进行性能优化？
- 美团新接入一个推荐场景，需要耗多少人力，算法与工程的人力配比是多少，算法与工程的分工是什么样的？
- 请问如何评估推荐系统？
- 请问 XGBoost、FM 是跑在 Spark 上的吗？FM 是自己实现的还是用的第三方库？
- 统计类特征受回流的日志是否正常，错误的日志得到错误的统计类特征，是否推荐算法不是实时的？

问：LR+FM 混合模型具体怎么做？

答：组合模型中，FM 的目的是用于做特征组合，LR 的目的是用于解决用户没有行为或用户行为较少时做推荐，思路有点像 Google 的 paper 《Wide & Deep Learning for Recommender Systems》，这篇 paper 里 LR 也是用作了 Memorization，网络是用于泛化，我们这里类似，LR 的目的是一样的，只不过我们是使用了 FM。

具体的做法是，特征包括：用户属性特征、用户行为特征、品类属性特征，所有用户行为特征都放到 FM 中做训练，其他特征都放到 LR 中训练，两个模型同时调同一个 LOSS。

问：数据清洗采用的是什么技术？用过 ETL 工具吗？

答：我们使用过 ETL，数据清洗我们主要是使用 Spark，ETL 一般用来做不需要任何业务处理的单纯过滤和整合，其他数据清洗基本都是用 Spark 完成的。

问：Spark 跟 Hadoop 的区别是什么？为什么要使用 Spark？

答： Spark 基于 Hadoop，只是计算的时候是在内存中计算，所以速度提升了许多。Spark 其实是用 map 和 reduce 操作封装了许多操作，同时对 Hive SQL 的执行做了优化，加入了 Streaming 和 MLlib。综上，Spark 不仅快而且因为封装了许多操作，所以写起来其实比 Hadoop 要更方便一些，同时支持很多机器学习模型，基于这些考虑，我们使用的是 Spark。

问：FM 中的 k 表示什么？

答： FM 中的 k 表示隐向量的维数，k 影响了两件事：

1. FM 模型的时间复杂度，FM 的时间复杂度是 kO(n)；
2. FM 模型的表达能力。

问：Spark 在贵公司的实际应用场景有哪些？Sparkmlib 是否比 skitlearn 和 TensorFlow 更实用一些？

答： Spark 的使用场景还是比较多的。目前，我们这里是使用 Spark 进行数据清洗、数据整合等一系列的数据作业开发，同时因为 Spark 非常适合大量的数据场景，所以我们还会使用 Spark 开发一些同步数据作业，把我们的推荐结果同步到 KV 数据库中。Spark 中的 MLlib 是有不少模型的，但是里边包含的模型大多都是方便并行执行的模型，所以很多模型是没有的，sklearn 是机器学习算法的集成，TensorFlow 是深度学习框架。是否实用的话，要看具体业务场景和需求。

问：Java Web 程序员如何转型推荐系统算法工程师？

答：其实我本身也是一个后端开发工程师，关于转型，可以给出一些建议。

1. 有数学基础的同学，可以从基础算法看起，比如线性回归、逻辑回归、SVM、随机森林之类的，具体需要的话，有英文基础的最好去看相关的论文，深入去熟悉一下这些算法的原理和细节，清楚不同算法之间的区别以及适用场景；
2. 没有数学基础的同学，最好先补一下数学基础，然后再去看第一个建议；
3. 最后一个建议就是，可以先用 Spark 跑一个非常基础的分类算法，比如跑一下辨别垃圾邮件，代码非常简单，然后看看结果，再去想想算法中是怎么实现的。

推荐两本书，《机器学习》（西瓜书）和《统计学习方法》。

问：请问为何不使用深度学习来推荐？如何做的特征筛选？FM 是 sparkmlib 自带的包还是自己实现的？以及用户画像是怎么实现的？

答： 几个问题答案分别如下：

1. 没有使用深度学习有一些历史原因，因为不是算法模型考虑上的问题，所以这里不做赘述。
2. 特征筛选我们目前是人工筛选的，也可以使用 XGBoost 这样的模型做特征筛选。
3. FM 是我们公司自实现的一个训练平台来训练的，FM 的作者使用 C++ 自实现了一个 libFM，GitHub: https://github.com/srendle/libfm 。
4. 用户画像不是我们做的，所以这里就不作过多回答了。

问：推荐系统有做 abtest 吗？

答：有的。我们 AB 的做法就是为新上线策略分一部分流量，同时也会分出相同的流量给一个对照组，做效果分析的时候用对照组和实验组的数据做对比。

问：请问模型是实时训练的吗？

答：不是实时训练的。目前我们是离线训练的。每天训练的是前一天产生的数据，训练结果会在有效行为的第二天展出。

问：设备 ID 到用户 ID 转换，扩展未登录场景，请问怎么扩展到未登录场景，比如 H5 未必存在获取到设备信息？怎么设计推荐评测系统进行验证，以便不断优化推荐算法？

答：设备 ID 到用户 ID 的转换主要针对的是设备上未登录用户曾经登录过的情况，在这种情况下，我们是会有一个映射关系的，根据这个映射关系进行设备 ID 到用户 ID 的转换。H5 获取不到设备信息这种不是我们这边需要考虑的点，我们只考虑已存在映射关系的数据转换。

PPT 第九页里主要讲的是增加覆盖率的一些措施。

1. 设备 ID 到用户 ID，这个已经讲解过了；
2. 扩大时间周期，这个具体是指，我们最初的模型数据只用了一周的数据，后期为了增加覆盖率，时间周期扩大到了两周、四周这样，所以覆盖了更多的用户；
3. 扩展行为数据就是增加一些新的行为数据，比如收藏数据；
4. 品类层级关系转换涉及到业务层面的东西，这里就不作详细介绍了。

问：美团推荐的召回是怎么做的，精排只用了 FM+LR 吗，重排是怎么做的？

答：我们的召回目前是全召回，因为我们这里不是搜索或者广告，搜索是需要根据用户输入的关键词根据相似性进行召回，广告是会根据用户群体进行召回，我们这里召回集不是非常大，所以是全召回的。

目前模型是使用了 FM+LR，重排序会根据业务需求进行重排。

问：美图推荐的用户画像有多少个特征，如果特征很大，怎么进行性能优化？

答：我们使用用户画像的特征不是非常多，大多数是行为特征 one-hot 编码后的特征。就目前而言，特征不是非常多。

特征如果很多的话，优化主要是训练平台层面的，我们是依托公司自开发的训练托管平台，据我所知，是基于 Parameter Server 研发的，分布式进行训练的，性能上来讲，还是非常不错的。

分类特征的编码方式会影响特征空间的大小，特征空间的大小以及 key 是否均匀在我们平台这里有可能会影响训练速度。所以如果特征很多做优化的话，是需要考虑这方面的。

问：美团新接入一个推荐场景，需要耗多少人力，算法与工程的人力配比是多少，算法与工程的分工是什么样的？

答：我觉得不同的业务场景是不同的。拿我们金刚区推荐来讲，金刚区接入用户个性化推荐的时候，算法建模、工程实现、数据作业开发这一系列都是我们自己完成的，人力的话大概是2个0.5人力和1个人完全做这件事。算法和工程配比这个我觉得也是看不同的业务场景和业务部门，不同的时期也不太一样。所以这里就不多作说明了。

问：请问如何评估推荐系统？

答：我们一般从两方面进行评估。

1. 业务指标；
2. 离线训练指标。

业务指标，做推荐这个事情的时候一定是有一个目标的，是要提升 CTR 还是提升其他指标，一般情况下，我们会把我们想要提升的东西当成我们的 label。推荐不管使用什么算法，其实都只是一种技术手段，最终都是要落地的，目的都是为公司带来收益，给用户带来更好的推荐结果，用户看到了自己想要的东西，业务也有自己的流量。所以，业务指标其实应该是最重要的一个衡量指标。

离线训练指标就是说，我们在训练模型的时候，是没办法确定上线后业务指标是什么样子的。所以只能先通过离线训练的一些指标来判断模型是否靠谱，比如用 AUC、LOSS。还有一些手段，比如随机选择一批用户，然后去看这批用户的预测结果和用户的特征，人工去看这个预测结果是否靠谱。用这个来调整模型的参数、特征等。在最终上线前，我们会看一下预测结果的数据集分布是否合理，是否符合预期，这也是我们评价一个模型的次要指标。

问：请问 XGBoost、FM 是跑在 Spark 上的吗？FM 是自己实现的还是用的第三方库？

答：不是跑在 Spark 上的，我们使用 Spark 只做了两件事。第一件事是数据作业的开发，一切数据相关的都在 Spark 上完成，包括数据清洗、数据整合、数据分析、数据处理。第二件事是同步作业，我们使用 Spark 进行数据同步，写入 KV 数据库。MLlib 是没有 FM 的。之前有人写过 SparkFM，GitHub 上可以找到，大家有需求的话可以自己去找一下。

问：统计类特征受回流的日志是否正常，错误的日志得到错误的统计类特征，是否推荐算法不是实时的？

答：我们的统计类特征是日志清洗后的 UV 数据，所以理论上来讲，应该是过滤掉了错误的日志。我们目前的推荐不是实时的，是离线训练的。而且我觉得，统计类特征包含实时数据无法描述的关联，在平衡资源的情况下，信息考虑的越多越好。